

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-273214

(43)Date of publication of application : 05.10.2001

(51)Int.Cl.

G06F 13/00

G06F 17/21

(21)Application number : 2000-083150

(71)Applicant : OKI SOFTWARE KK
OKI ELECTRIC IND CO LTD
NTT ME CORP

(22)Date of filing : 24.03.2000

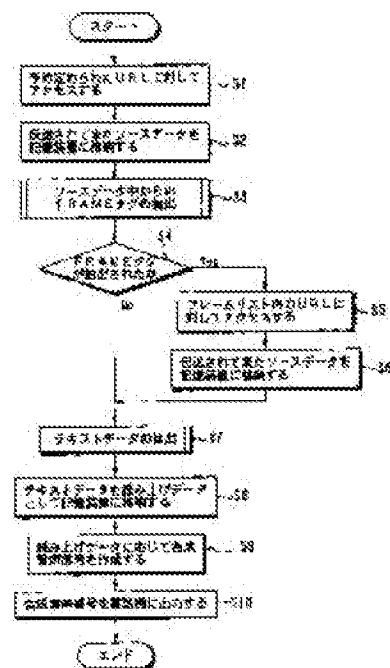
(72)Inventor : NISHIMURA HITOSHI
YAKIDA KAZUHIKO
MATSUSHITA ARIYUKI
YAMAGUCHI YUICHIRO
ITO SHINICHI
NAGAI TOMOYASU
OTA TSUYOSHI
TAMURA MASARU

(54) WEB PAGE DECODING SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a Web page decoding system capable of exactly extracting a text part in respect to the document of an HTML containing a tag having the description of a uniform resource locator(URL).

SOLUTION: Basic source data comprising a Web page are extracted from a storage area designated by the prescribed URL, and written in a storage means and when the existence of the tag containing the description of the URL is detected out of the basic source data, the URL in that prescribed tag is detected. Then, source data are extracted from the storage area designated by that detected URL, and written in the storage means and the text part is extracted from all the source data stored in the storage means.



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2001-273214
(P2001-273214A)

(43) 公開日 平成13年10月5日 (2001.10.5)

(51) Int.Cl. ⁷	識別記号	F I	テームコード* (参考)
G 0 6 F 13/00	3 5 4	G 0 6 F 13/00	3 5 4 D 5 B 0 0 9
17/21	5 0 1	17/21	5 0 1 T 5 B 0 8 9
	5 6 8		5 6 8 A
	5 9 6		5 9 6 A

審査請求 未請求 請求項の数 6 O L (全 6 頁)

(21) 出願番号 特願2000-83150(P2000-83150)

(22) 出願日 平成12年3月24日 (2000.3.24)

(71) 出願人 591051645

沖ソフトウェア株式会社
東京都板橋区舟渡1丁目12番11号

(71) 出願人 000000295

沖電気工業株式会社
東京都港区虎ノ門1丁目7番12号

(71) 出願人 596094692

株式会社エヌ・ティ・ティ エムイー
東京都千代田区大手町二丁目2番2号

(74) 代理人 100079119

弁理士 藤村 元彦

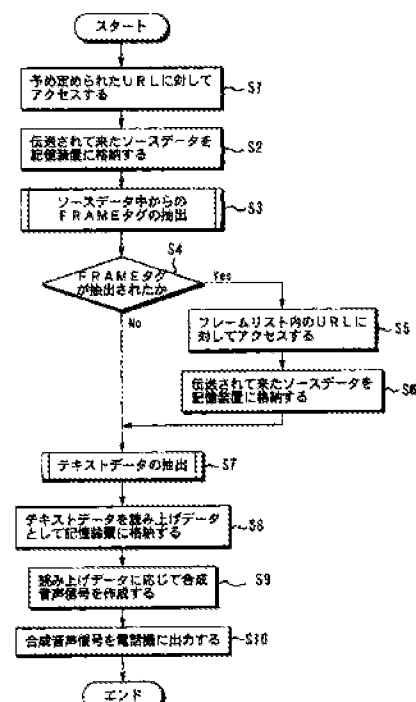
最終頁に続く

(54) 【発明の名称】 ウェブページ解説システム

(57) 【要約】

【課題】 URL (ユニホームリソースロケータ) の記述を有するタグを含む HTML の文書に対してテキスト部分を正確に抽出することができる Web (ウェブ) ページ解説システムを提供する。

【解決手段】 Web ページを構成する基本のソースデータを所定の URL で指定された記憶領域から取り出して記憶手段に書き込み、基本のソースデータ中から URL の記述箇所を含む所定のタグの存在を検出した場合には、その所定のタグ中の URL を検出し、その検出した URL で指定された記憶領域からソースデータを取り出して記憶手段に書き込み、記憶手段に記憶されたソースデータの全てからテキスト部分を抽出する。



【特許請求の範囲】

【請求項1】 Web（ウェブ）ページを構成するHTML文書のテキスト部分を解読するWebページ解読システムであって、

前記Webページを構成する基本のソースデータを所定のURL（ユニホームリソースロケータ）で指定された記憶領域から取り出して記憶手段に書き込む手段と、

前記基本のソースデータ中からURLの記述箇所を含む所定のタグの存在を検出するタグ検出手段と、

前記所定のタグの存在が検出された場合にはその所定のタグ中のURLを検出するURL検出手段と、

前記URL検出手段によって検出されたURLで指定された記憶領域からソースデータを取り出して前記記憶手段に書き込む手段と、

前記記憶手段に記憶されたソースデータの全てからテキスト部分を抽出するテキスト抽出手段と、を備えたことを特徴とするウェブページ解読システム。

【請求項2】 前記所定のタグはフレームタグであることを特徴とする請求項1記載のウェブページ解読システム。

【請求項3】 前記URL検出手段は、前記フレームタグ中のURLとして<FRAME SRC="URL">の構文中のURLを検出することを特徴とする請求項1又は2記載のウェブページ解読システム。

【請求項4】 前記テキスト抽出手段は、前記記憶手段に記憶されたソースデータ中の<>で囲まれた部分以外の部分をテキスト部分として抽出することを特徴とする請求項1記載のウェブページ解読システム。

【請求項5】 前記テキスト抽出手段によって抽出されたテキスト部分に対応して音声信号を作成して出力する音声出力手段を更に備えたことを特徴とする請求項1記載のウェブページ解読システム。

【請求項6】 前記音声出力手段の出力音声信号は公衆電話回線を介して電話機に供給されることを特徴とする請求項1記載のウェブページ解読システム。

【発明の詳細な説明】

【0001】

【発明が属する技術分野】 本発明は、Web（ウェブ）ページのテキスト部分を解読するWebページ解読システムに関する。

【0002】

【従来の技術】 インターネットの情報サービスの1つであるWWW(World Wide Web)は、HTML(Hyper Text Markup Language)という言語で記述されたHTMLファイルとそのファイルの保存位置の識別子であるURL(Uniform Resource Locator)とを用いてインターネットを介して文字、映像、音声等のマルチメディア情報を参照することができるものである。HTMLファイルをWWWブラウザと呼ばれる閲覧ソフトウェアによって処理することによりディスプレイ画面上に形成されるものがW

ebページである。WWWの情報を提供する側、すなわちWWWサーバはHTMLファイルをURLで関連付けて保存しており、サーバアプリケーションに従って動作する。情報を提供される側、すなわちクライアント（クライアントコンピュータ）ではWWWブラウザを用いて所望のURLからHTMLファイルを含むソースデータ（例えば、画像ファイルや音声ファイル）にインターネットを介してアクセスしてソース中のファイルによって形成されたWebページをディスプレイ画面上で参照することができる。

【0003】 HTMLの文書は、通常、テキスト文書からなるテキスト部とタグによって形成される表示情報部とから形成される。タグは<>を一对とする記号であり、タグを用いてHTMLの構文が例えば、<HTML>～</HTML>の如く形成される。タグ<>で囲まれた部分にはWebページに表示されるテキスト部の文字の大きさ、フォントの種類、その文字色、Webページの背景色、画像ファイル名、画像位置等の様々な表示情報が示される。

【0004】 このようにHTMLの文書においてはタグ<>で囲まれた部分は表示される部分ではなく、文書を表示するための制御情報であるので、タグ<>で囲まれた部分を除くと、通常、単なるテキスト文書となるのが普通である。一方、WWWブラウザに表示されるWebページの文書を読み上げるシステムがインターネット上には形成されることがある。これは、クライアントの端末がコンピュータではなく、例えば、公衆回線に接続された電話機である場合にWebページの文書を読み上げて音声信号として電話機に送出するためである。Webページ読上としては読み上げ対象のHTMLファイル中からタグ<>で囲まれた部分を除くテキストデータ部分を抽出し、そのテキストデータ部分の文字コードに対応した音声データを合成して一連の音声信号として出力することが行われる。

【0005】

【発明が解決しようとする課題】 HTMLには、Webページ上にフレームを形成するための構文として、例えば、<FRAMESET>～</FRAMESET>があり、これを用いたWebページでは分割された画面が得られる。その構文が記述されたHTMLファイルからは分割画面毎に別のHTMLファイルが更に呼び出されて文書が表示される。すなわち、<FRAME SRC="URL">の如きタグにより分割画面毎にURL（HTMLファイル名を含む）が指定され、その指定されたURLの領域に存在するHTMLファイルの内容が表示される。

【0006】 しかしながら、従来、このようなフレームタグのようにURLの記述を有するタグを含むHTMLの文書に対してはテキスト部分を正確に抽出することができないという問題点があった。そこで、本発明の目的

は、URLの記述を有するタグを含むHTMLの文書に対してテキスト部分を正確に抽出することができるWebページ解読システムを提供することである。

【0007】

【課題を解決するための手段】本発明のWebページ解読システムは、Webページを構成するHTML文書のテキスト部分を解読するWebページ解読システムであって、Webページを構成する基本のソースデータを所定のURL(ユニホームリソースロケータ)で指定された記憶領域から取り出して記憶手段に書き込む手段と、基本のソースデータ中からURLの記述箇所を含む所定のタグの存在を検出するタグ検出手段と、所定のタグの存在が検出された場合にはその所定のタグ中のURLを検出するURL検出手段と、URL検出手段によって検出されたURLで指定された記憶領域からソースデータを取り出して記憶手段に書き込む手段と、記憶手段に記憶されたソースデータの全てからテキスト部分を抽出するテキスト抽出手段と、を備えたことを特徴としている。この構成より、基本のソースデータがフレームタグを含むHTMLファイルの場合には、そのフレームタグ内に記述されたURLの領域に格納されているHTMLファイルのテキスト部分も抽出することができる。

【0008】

【発明の実施の形態】以下、本発明の実施例を図面を参照しつつ詳細に説明する。図1は本発明のによるWebページ解読システムの構成を示している。このシステムにおいては、WWWサーバ1は情報サービスとしてWWWを提供するサーバであり、HTMLファイルをURLで関連付けて保存しており、また、画像ファイルや音声ファイルも保存している。WWWサーバ1はインターネット回線網2に接続されている。

【0009】インターネット回線網2にはCTI(Computer Telephony Integration)サーバ3が接続されている。CTIサーバ3は公衆電話回線網4にも接続されている。公衆電話回線網4には複数の電話機が実際には接続されているが、ここでは1つの電話機5を示している。電話機は一般加入電話機、公衆電話機及び携帯電話機のいずれであっても良い。なお、公衆電話回線網4には中継局、基地局等の電話回線接続のための局が存在するが、図には示していない。

【0010】CTIサーバ3はWebページの読み上げを電話機5を含む電話機のユーザに提供するサーバである。CTIサーバ3には、Webページ取得部31と、テキスト抽出部32と、テキスト読み上げ部33とが備えられている。Webページ取得部31はWWWサーバ1にアクセスし、Webページのソースデータを取得する。テキスト抽出部32はWebページ取得部31によって取得されたソースデータを解析し、テキスト部分を抽出する。テキスト読み上げ部33はテキストの文字コードに応じて合成音声信号を作成し、その合成音声信号

を公衆電話回線網4を利用して電話機に対して出力する。Webページ取得部31、テキスト抽出部32及びテキスト読み上げ部33はCTIサーバ3のプロセッサ(図示せず)の後述の如き動作によって形成される。

【0011】また、CTIサーバ3は内部にハードディスク等の記憶装置35を有しており、後述するように、ソースデータ等の各種データが記憶装置35には記憶される。WWWサーバ1及びCTIサーバ3各々のインターネット回線網2を利用した通信においては通信プロトコルとしてTCP/IPが用いられ、WWWサーバ1及びCTIサーバ3にはIPアドレスが各々割り当てられている。更に、WWWのプロトコルとしてはHTTPが使用される。また、図示していないが、WWWサーバ1及びCTIサーバ3はルータを介してインターネット回線網2には接続されている。

【0012】次に、かかるWebページ解読システムの動作について説明する。ユーザが電話機5からCTIサーバ3へ電話をかけ、電話機5とCTIサーバ3との間の通話状態が確立すると、CTIサーバ3は図2に示すように、先ず、予め定められたURLで指定される領域のWebページのソースデータを取得するために、そのURLに対してアクセスを行う(ステップS1)。このURLがWWWサーバ1内にあるとすると、WWWサーバ1はURLで指定される領域のHTMLファイル等のファイルからなるソースデータを読み出してCTIサーバ3に対して送信する。そのソースデータはWebページを構成する基本となるソースデータである。送信されたソースデータはインターネット回線網2を介してCTIサーバ3に供給される。

【0013】CTIサーバ3はWWWサーバ1から送られて来たソースデータを記憶装置35に格納し(ステップS2)、その格納したソースデータ中からFRAME(フレーム)タグを抽出する(ステップS3)。このFRAMEタグの抽出動作については後述するが、FRAMEタグ中に含まれるURLがフレームリストとして記憶装置35に書き込まれる。

【0014】CTIサーバ3はステップS3の実行の結果として、FRAMEタグの抽出が行われたか否かを判別する(ステップS4)。ステップS4にてFRAMEタグの抽出が実際に行われなかった場合には、後述のステップS7に進む。一方、FRAMEタグの抽出が実際に行われた場合には、フレームリストのURLで指定される領域のWebページのソースデータを取得するために、そのフレームリストのURLに対してアクセスを行い(ステップS5)、WWWサーバ1から送られて来たソースデータを記憶装置35に格納する(ステップS6)。ステップS5のアクセスに対するWWWサーバ1の動作はステップS1のアクセスの場合と同様である。ステップS6の実行後はステップS7に進む。

【0015】CTIサーバ3はステップS7において記

憶装置35に格納されたソースデータからタブ<>で囲まれた部分以外のテキスト部分を抽出し、その抽出テキストデータを読み上げデータとして記憶装置35に書き込む(ステップS8)。その後、読み上げデータに基づいて合成音声信号を作成し(ステップS9)、その合成音声信号を電話機5に対して出力する(ステップS10)。記憶装置35に書き込まれた読み上げデータは複数の文字コードからなるテキストデータであるので、その文字コード各々又は単語単位の文字コード群に対応する音声データを記憶装置35から検索して得て、それら

10 音声データを合成して連続する合成音声信号を作成する。合成音声信号は公衆電話回線網4を介して電話機5に供給され、電話機5の受話器から読み上げ音出力される。なお、記憶装置35には文字コードと音声データとの関係を示すデータテーブルが予め記憶されている。
【0016】次に、上記のステップS3におけるソースデータ中からのFRAMEタグ抽出動作について図3のフローチャートを参照しつつ説明する。CTIサーバ3は、記憶装置35に記憶されたソースデータ中から文字列<FRAME SRCを検索し(ステップS11)、文字列<FRAME SRCがソースデータ中に存在する

20 かどうかを判別する(ステップS12)。すなわち、記憶装置35に書き込まれたソースデータ中にはHTMLファイルが含まれ、そのHTMLファイルが示す文書でフレーム設定が行われているかどうかを判別される。文字列<FRAME SRCが存在するならば、<FRAME SRC="URL">の構文が存在するので、読み取り位置をその次の文字=の位置まで移動し(ステップS13)、更に、その後の" "で囲まれた文字列、すなわちURLをソースデータから読み取り、そのURLを記憶装置35に形成されたフレームリストに書き込む(ステップS14)。よって、フレームリストにはフレーム内に含まれるHTML文書の存在位置を示すURLが書き込まれる。ステップS14の実行後、ソースデータの全てのファイルから文字列<FRAME SRCの検索が終了したかどうかを判別し(ステップS15)、その検索が終了していない場合にはステップS11に戻り、上記のステップ動作を繰り返す。

【0017】次いで、上記したステップS7におけるテキストデータの抽出動作について図4のフローチャートを参照しつつ説明する。CTIサーバ3は、記憶装置35に格納されたソースデータのうちの1つのファイルの先頭から順に1文字分の文字コードを取得し(ステップS21)、その文字コードが文字<を示すかどうかを判別する(ステップS22)。取得した文字コードが文字<を示す場合にはタグフラグF_{TM}を1に等しくさせる(ステップS23)。取得した文字コードが文字<を示さない場合にはタグフラグF_{TM}が1に等しいかどうかを判別する(ステップS24)。タグフラグF_{TM}はHTMLファイルにおいて<>で囲まれた部分において1に

設定され、それ以外の部分において0に設定されるフラグであり、その初期値は0である。ステップS23の実行後もステップS24の判別は実行される。

【0018】ステップS24の判別の結果、タグフラグF_{TM}が1に等しくされている場合には、ステップS21で取得した文字コードが文字>を示すかどうかを判別する(ステップS25)。取得した文字コードが文字>を示す場合にはタグの終了であるので、タグフラグF_{TM}を0に等しくさせる(ステップS26)。一方、ステップS24の判別の結果、タグフラグF_{TM}が0に等しくされている場合には、<>で囲まれたタグ部分以外のテキスト部分であるので、取得した文字コードを読み上げデータに含ませるように記憶装置35に格納する(ステップS27)。

【0019】ステップS26又はS27の実行後はソースデータの全てから1文字分の文字コードの取得が終了したかどうかを判別し(ステップS28)、その取得が終了していない場合にはステップS21に戻り、上記のステップ動作を繰り返す。よって、かかるシステムによれば、基本のソースデータがFRAMEタグを含むHTMLファイルの場合にFRAMEタグ内に記述されたURLの領域に格納されているHTMLファイルのテキスト部分も抽出することができるので、WWWブラウザに表示されるWebページのテキスト部分を余すことなく読み上げることができる。

【0020】上記した実施例においては、フレームタグが使用された場合について説明したが、フレームタグ以外のURLの記述を有するタグにも本発明を適用することができる。また、JavaScript等のスクリプト言語を含むHTMLファイルからテキスト部分を抽出して読み上げる場合にも本発明を適用することができる。HTMLでは、機能を拡張するためにWebページ上でJavaScript等のスクリプト言語を実行できるようにするタグも用意されている。例えば、<SCRIPT LANGUAGE="JavaScript">~</SCRIPT>のような構文で形成される。よって、上記したように<SCRIPT LANGUAGE="JavaScript">~</SCRIPT>の範囲の部分を見捨てそれ以外のテキスト部分を抽出するのである。

【0021】更に、本発明のシステムはHTMLを用いたファイルの場合に限らず、HTMLを拡張させた言語を用いたファイルについても適用することができる。なお、日本ではWWWブラウザにて閲覧できるページを全てホームページと称しているが、ホームページは本来、1情報群を構成する複数のWebページのうちの基本ページであるので、ここでは誤解を招かないようにWebページと記載した。

【0022】

【発明の効果】以上の如く、本発明のWebページ解説

システムにおいては、URLの記述を有するタグを含むHTMLの文書に対してテキスト部分を正確に抽出することができる。よって、Webページを読み上げる際にはWebページのテキスト部分を余すことなく読み上げることができる。

【図面の簡単な説明】

【図1】本発明によるWebページ解読システムの構成を示すブロック図である。

【図2】図1のシステム中のCTIサーバの動作を示すフローチャートである。

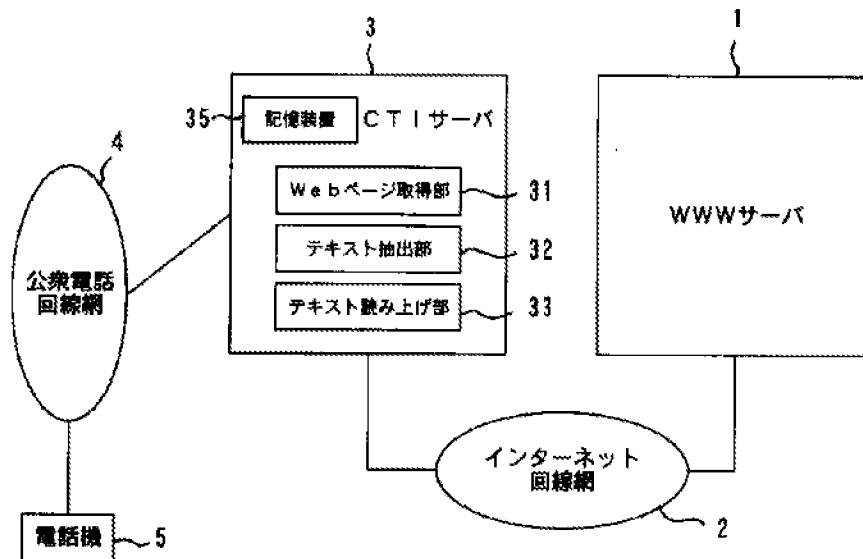
*【図3】ソースデータ中からのFRAMEタグ抽出動作を示すフローチャートである。

【図4】テキストデータの抽出動作を示すフローチャートである。

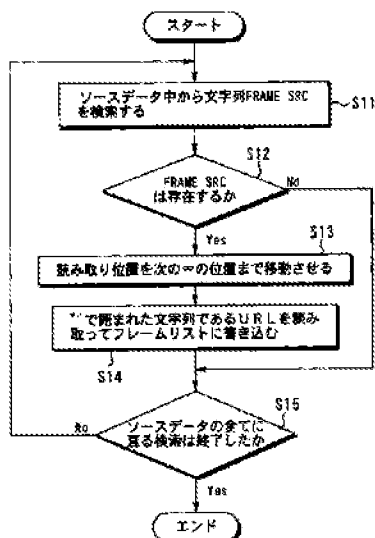
【符号の説明】

- 1 WWWサーバ
2 インターネット回線網
3 CTIサーバ
4 公衆電話回線網
5 電話機

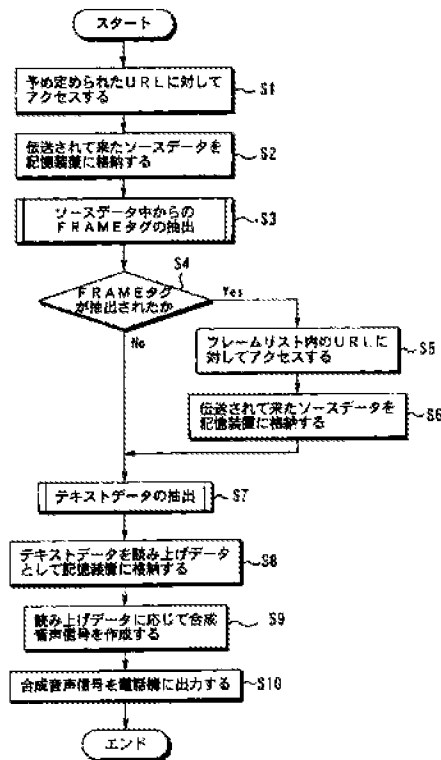
【図1】



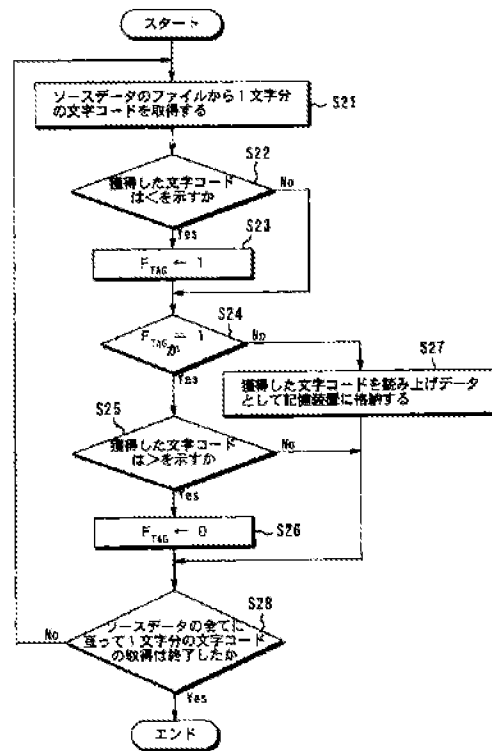
【図3】



【図2】



【図4】



フロントページの続き

- (72)発明者 西村 仁司
東京都板橋区舟渡1丁目12番11号 沖ソフトウェア株式会社内
- (72)発明者 八木田 一彦
東京都板橋区舟渡1丁目12番11号 沖ソフトウェア株式会社内
- (72)発明者 松下 有亨
東京都板橋区舟渡1丁目12番11号 沖ソフトウェア株式会社内
- (72)発明者 山口 雄一郎
東京都港区虎ノ門1丁目7番12号 沖電気工業株式会社内
- (72)発明者 伊藤 慎一
東京都港区虎ノ門1丁目7番12号 沖電気工業株式会社内

- (72)発明者 永井 友康
東京都千代田区大手町2-2-2 アーバンネット大手町ビル 株式会社エヌ・ティ・ティエムイー内
- (72)発明者 大田 剛志
東京都千代田区大手町2-2-2 アーバンネット大手町ビル 株式会社エヌ・ティ・ティエムイー内
- (72)発明者 田村 賢
東京都千代田区大手町2-2-2 アーバンネット大手町ビル 株式会社エヌ・ティ・ティエムイー内
- Fターム(参考) 5B009 QA11 RD03 SA03 SA14 TA11
VA02 VC01
5B089 GA11 GB03 HA01 JA22 JB02
KA04 KB07 KC53 KC59 LB13